DOCUMENT RESUME

ED 183 568 TM 010 018

AUTHOR Yen, Wendy M.
TITLE Using Simulation Results When Choosing a Latent-Trait Model.
PUB DATE 25 Sep 79
NOTE 24p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Correlation; Goodness of Fit; *Item Analysis; Junior High Schools; *Mathematical Models; Multiple Choice Tests; Psychometrics; *Simulation; Statistical Analysis; Statistical Bias; Test Items; *Test Results
IDENTIFIERS Comprehensive Tests of Basic Skills; *Latent Trait Models; Rasch Model; *Test Analysis; Three Parameter Model; Two Parameter Model

ABSTRACT

Three test-analysis models were used to analyze three types of simulated test score data plus the results of eight achievement tests. Chi-square goodness-of-fit statistics were used to evaluate the appropriateness of the models to the four kinds of data. Data were generated to simulate the responses of 1,000 students to 36 pseudo-items by procedures assuming a three-parameter, a two-parameter, and a one-parameter model. These three sets of pseudo-responses plus the data from eight achievement tests given to seventh and eighth grade children were analyzed by the three methods. The three-parameter logistic model provided the best fit in most instances. Ability estimates based on the two-parameter and the three-parameter models had a curvilinear relationship. The two-parameter estimation of item discriminating power was unacceptable, being strongly affected by the item's difficulty level. Data for the eight actual multiple-choice tests were most consistent with the three-parameter model. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED183568

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Using Simulation Results When Choosing A Latent-Trait Model

Wendy M. Yen

CTB/McGraw-Hill

Running head: Choosing a Latent-Trait Model

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Wendy M. Yen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Address correspondence concerning this paper to:

Wendy M. Yen

CTB/McGraw-Hill

Del Monte Research Park

Monterey, CA 93940

Manuscript submitted for publication: September 25, 1979

TM010 014

Abstract

Simulated data were generated for 36 pseudo-items and 1000 pseudo-examinees using three-, two-, and one-parameter logistic latent-trait models. The data were analyzed using three-, two-, and one-parameter models. Between model comparisons were made of chi-square goodness-of-fit statistics, trait estimates, and item-parameter estimates. The three generating models produced clearly different patterns of results. The simulation results were compared to results for real data involving seventh- and eighth-grade students' performance on eight achievement tests. The achievement results appeared most similar to the simulation results based on data generated with the three-parameter model.

There are three latent-trait models that are in common use. The three-parameter (3-PAR) model estimates item difficulties, item discriminating powers, and item lower asymptotes. The two-parameter (2-PAR) model estimates item difficulties and discriminating powers but assumes that all lower asymptotes are zero. The one-parameter (1-PAR) or Rasch model estimates item difficulties but assumes that all item discriminating powers are constant and all lower asymptotes are zero. (See Allen & Yen, 1979, for a further discussion of these models.) When items have a multiple-choice format and it is possible for examinees of low ability to get an item correct through lucky guessing, the 2-PAR and 1-PAR models appear to be inappropriate. However, in striving for simplicity the researcher may be tempted to use the 2-PAR or 1-PAR models and hope that the inaccuracies that are introduced are unimportant. The researcher may turn to a statistical goodness-of-fit test to gauge the degree of these inaccuracies, even though it is not always clear how to interpret differences in levels of fit. This paper presents the results of a simulation study that can aid in the interpretation of a chi-square goodness-of-fit measure and that offer additional, qualitative methods of judging the suitability of the various latent-trait models. Real data examples are given of how the simulation results can be used in helping choose an appropriate latent-trait model.

Method

Item responses were generated to simulate the performance of students from two adjacent grades. 500 trait values (θ) from each of two normal distributions were used. One θ distribution had a

mean of - .18 and a standard deviation of .95; the other θ distribution had a mean of .16 and a standard deviation of 1.02. These two distributions were pooled to produce one distribution of 1000 latent traits. Item discriminating powers, a, item difficulties, b, and item lower asymptotes, c, had unimodal, somewhat platykurtic distributions similar to those found with real data. The means and standard deviations of these distributions were $\bar{a} = 1.15$, $s_a = .40$, $\bar{b} = -.05$, $s_b = 1.15$, $\bar{c} = .24$, and $s_c = .04$. Parameter sets for 36 items were created by randomly choosing without replacement from the parameter distributions an a, b, and c for each item. The resulting correlations taken over items of the parameters were $r_{ab} = .20$, $r_{ac} = -.23$, and $r_{bc} = .08$.

For the three-parameter model the probability of pseudo-examinee k passing pseudo-item [1] i was generated as

Insert Equation 1 here

For each (i,k) pair a random number from 0 to 1 was generated from a uniform distribution using IMSL subroutine GGUBS (IMSL, 1979). If the random number was less than $p_i(\theta_k)$, examinee k was said to pass item i; otherwise, examinee k failed item i. This item response generation procedure resulted in pass/fail response arrays based on the 3-PAR model for 36 items for 1000 examinees.

[1] The "pseudo" will now be dropped for the sake of simplicity, and it should be obvious that for the simulation data the examinees and items were not real.

Items responses were also generated using the 2-PAR and 1-PAR models. For the 2-PAR model c_i was set equal to 0 for all items, and then the pass/fail responses were generated for the 36 items and 1000 examinees using Equation 1. For the 1-PAR model, c_i was set equal to 0 and a_i was set equal to 1/1.7 for all items, and then the pass/fail responses were generated. Thus, there were three sets of item responses, one set generated with the 1-PAR model, another set generated with the 2-PAR model, and a third set generated with 3-PAR model.

Each of the three sets of item responses was analyzed with each of three estimating models, producing nine analyses. The item responses were analyzed with the LOGIST computer program (Wood, Wingersky, & Lord, 1976) to produce estimated trait values ($\hat{\theta}$) and item parameters ($\hat{a}$, $\hat{b}$, $\hat{c}$). When default options were used, LOGIST automatically estimated traits and parameters using the 3-PAR model. To estimate traits and parameters using the 2-PAR model, LOGIST was constrained to hold $\hat{c} = 0$ for all items. To estimate traits and parameters using the 1-PAR model, LOGIST was constrained to hold $\hat{c} = 0$ and $\hat{a} = 1/1.7$ for all items.

A chi-square goodness-of-fit test was calculated for each of the nine combinations of generating models and estimating models. To calculate the chi-square for a given combination, the 1000 examinees were rank ordered on the basis of their trait estimates and then divided into ten cells with approximately equal numbers of examinees in each cell. The chi-square for an item was calculated as

---

Insert Equation 2 here

---

where N_j was the number of examinees in cell j, O_{ij} was the observed proportion of examinees in cell j that passed item i, and E_{ij} was the proportion of examinees in cell j expected to pass item i.

Insert Equation 3 here

where $\hat{p}_i(\hat{\theta}_k)$ was the item-characteristic function evaluated using the trait estimated for examinee k and the item parameters estimated for item i. This chi-square statistic had 10 - 3 = 7 degrees of freedom when the 3-PAR model was used in the estimation, 10 - 2 = 8 degrees of freedom when the 2-PAR model was used in the estimation, and 10 - 1 = 9 degrees of freedom when the 1-PAR model was used in the estimation.

To examine the extent to which the simulation results were helpful in understanding real data, eight tests from the Comprehensive Tests of Basic Skills, Form S, Level 3 (1973) were analyzed. The tests were Reading Vocabulary (40 items), Reading Comprehension (45 items), Spelling (30 items), Language Mechanics (20 items), Language Expression (35 items), Mathematics Computation (48 items), Mathematics Concepts (25 items), and Mathematics Applications (25 items). These are multiple-choice tests with four answer choices per item. Data for 1000 seventh- and eighth-grade students were analyzed with the three latent-trait models.

Results

Simulation Data

Table 1 contains the means taken over the 36 items of the chi-

Insert Table 1 about here

squares for the simulation data. The columns of chi-square values had very similar patterns when the data were generated by the 3-PAR model and by the 2-PAR model. When the data were generated by the 3-PAR model, the 2-PAR estimating model did a surprisingly good job of fitting the data. If we were analyzing a set of real data with all three models, we probably would have a difficult time deciding on the basis of the pattern of chi-squares whether the 3-PAR or 2-PAR model was more appropriate. It would be clear that the 1-PAR model was appropriate when the mean chi-squares had similar values for all three estimating models. Keep in mind that real data tend to produce chi-squares that are greater than those in Table 1 because real data are generated by processes more complicated than the process described in Equation 1.

When the model used for estimating traits and item parameters matched the model used for generating a set of data, the estimated traits, item difficulties, and discriminating powers had strong linear relationships with the traits, item difficulties, and discriminating powers in the generating model. In such a situation the lowest correlation of estimated values with the generating values was .89, which was found for the 3-PAR discriminating powers. The square root of the mean squared difference between estimated and generating lower asymptotes for the 3-PAR model was .04. In other words, when an estimating model was appropriate, it tended to accurately recover the traits and parameters of the generating model. Because of this finding, results are presented only for estimated

traits and item parameters. Results involving the generating traits and parameters can be inferred from the results involving the estimated traits and parameters for the estimating model that matches the generating model.

The relationships between latent traits estimated by the three models are contained in Figure 1. Note that because the selection

---

Insert Figure 1 about here

---

of the scale in the estimation process is arbitrary, numberical labels of the axes of the plots are not important. When data were generated with the 1-PAR or 2-PAR models, every pair of estimated latent traits was essentially linearly related except at the very lowest trait values. More of a distinction was made among examinees with very low scores when traits were estimated with the 3-PAR or 2-PAR models that when they were estimated with the 1-PAR model. When data were generated with the 3-PAR model, pronounced nonlinear relationships between the estimated trait values were created. The 2-PAR estimating model tended to stretch out the trait scale (i.e., discriminate more among examinees) at the high end more than the 3-PAR estimating model. The 3-PAR estimating model tended to stretch out the trait scale more at the low end than the 1-PAR estimating model, and the 2-PAR estimating model tended to stretch out the trait scale more at the upper end than the 1-PAR estimating model. Notice that an examination of the correlations would not have revealed these patterns of relationships between the various estimated trait values.

Figure 2 displays the relationships between the item difficulties

---

Insert Figure 2 about here

---

estimated by the three models. The pattern of these relationships parallels the pattern found for the traits, but because there are fewer items than examinees, the patterns are more difficult to see with the items than with the trait values.

When data were generated with the 1-PAR or 2-PAR models with $c = 0$, the 3-PAR estimating model produced $c = 0$ for all items. When data were generated with the 1-PAR model with constant item discriminating powers, both the 2-PAR and 3-PAR estimating models produced constant item discriminating powers. For data generated with the 2-PAR and 3-PAR models, the relationships between the estimated item discriminating powers and the relationships between item discriminating powers and item difficulties are contained in Figure 3.

Insert Figure 3 about here

The relationships between the 3-PAR and 2-PAR estimated item discriminating powers were dramatically different depending on whether the data were generated with the 3-PAR or 2-PAR model. When the 3-PAR model generated the data, the 3-PAR and 2-PAR estimated discriminating powers had a weak relationship. When the 2-PAR model generated the data, the 3-PAR and 2-PAR estimated discriminating powers had an extremely strong linear relationship.

When the 3-PAR model was used for estimation, there was a low

positive correlation between $\hat{a}$ and $\hat{b}$ regardless of which model generated the data. This low positive correlation was consistent with the low positive correlation between a and b for the item parameters used in the data generation. When the 2-PAR model was used for estimation, there was a substantial difference in the correlation between $\hat{a}$ and $\hat{b}$ found when the 2-PAR model generated the data versus when the 3-PAR model generated the data. If the 2-PAR model generated the data, the low positive correlation between $\hat{a}$ and $\hat{b}$ appeared. However, if the 3-PAR model generated the data, use of the 2-PAR model for estimation produced a strong negative correlation between $\hat{a}$ and $\hat{b}$.

Real Data

The mean item chi-squares for the three latent-trait models for each of the eight CTBS tests are contained in Table 2. These

---

Insert Table 2 about here

---

chi-squares have patterns similar to those in Table 1 for the simulation data generated with the 3-PAR and 2-PAR models. The estimated trait values and item parameters for each CTBS test were examined, and their relationships appeared very similar to the relationships found for the simulation data generated with the 3-PAR model. For example, Table 3 contains the correlations

---

Insert Table 3 about here

---

between the estimated item difficulties and discriminating powers for the 3-PAR and 2-PAR models. These correlations are similar in pattern to the correlations found with simulation data generated with the 3-PAR model. All the results for Reading Vocabulary, which were typical of the results for the other tests, are contained in Figure 4. The $\hat{\theta}$ plots in the first column of Figure 4 can be compared

---

Insert Figure 4 about here

---

with each of the columns of plots in Figure 1; the Reading Vocabulary estimated traits appear most similar to the simulation traits based on data generated with the 3-PAR model. The $\hat{b}$ plots in the second column of Figure 4 can be compared with each of the columns of plots in Figure 2; it is not clear which data generation model is most consistent with the Reading Vocabulary item difficulties. The column of $\hat{a}$ plots in Figure 4 can be compared to the two columns of plots in Figure 3; the estimated item discriminating powers for the Reading Vocabulary data clearly are most consistent with the results for the simulation data generated with the 3-PAR model.

Discussion

The most surprising simulation results appeared when data that were generated with the 3-PAR model were analyzed with the 2-PAR model. (a) The 2-PAR model fit the data almost as well as the 3-PAR model fit the data (Table 1). (b) The traits (and item difficulties) as estimated by the 2-PAR and 3-PAR models were nonlinearly related,

with the 2-PAR estimated traits more stretched out at the high end than the 3-PAR traits (Figure 1). (c) The 2-PAR and 3-PAR estimated item discriminating powers had a weak relationship (Figure 3). (d) The 2-PAR estimated item difficulties and discriminating powers displayed a strong negative correlation (Figure 3). The explanation of these results centers around the fact that the 2-PAR model with c = o attempts to fit or explain item-characteristic functions with nonzero asymptotes. It is most awkward for the 2-PAR model to explain a nonzero asymptote when an item is difficult and has a moderate to high discriminating power. When an item is easy, there is little data available to establish the item's lower asymptote. When an item has a low discriminating power, the nonzero asymptote can be fairly well explained solely on the basis of a low discriminating power. The 2-PAR model deals with the difficult, highly discriminating items with nonzero asymptotes by stretching out the trait (and difficulty) scale at the high end. This stretching decreases the slope of the item-characteristic curves of difficult items, and the resulting item-characteristic curves can be explained with low discriminating powers. Thus, although the 2-PAR model is almost as accurate as the 3-PAR model in explaining pass/fail response arrays generated by a 3-PAR model, the 2-PAR model produces traits and item parameters that are very different from the traits and parameters used in generating the data. The 2-PAR model essentially preserves the rank order of examinees' trait values and item difficulties, but the 2-PAR estimation of discriminating powers is unacceptable for 3-PAR data. It is unacceptable because it is sample dependent. An item's discriminating power is strongly affected by the item's difficulty level relative

to the ability level of the group of examinees; items will tend to have systematically lower discriminating powers if they happen to be calibrated using a group of examinees with a relatively low ability level than if they happen to be calibrated with relatively able examinees. This sample dependency means that the predictions of the 2-PAR model can be accurate for the sample in which the 2-PAR item parameters are estimated, but the predictions can be inaccurate in cross-validations involving samples of examinees that differ systematically from the sample in which the parameters were estimated. Keep in mind that these inadequacies in the 2-PAR estimating model occur when the data are generated with the 3-PAR model; the 2-PAR estimating model works well with data generated with 2-PAR or 1-PAR models.

These simulation results provide information that can be useful in choosing which latent-trait model is most appropriate for a set of data. A chi-square goodness-of-fit test does not give complete information about the suitability of a model. When searching for the most suitable model, a researcher can analyze a set of data with all three models and compare the results with the simulation results in Figures 1 to 3. When this comparison was made for the eight CTBS tests, it was clear that the data were most consistent with the 3-PAR model.

Of course, there are other considerations that are important in choosing a latent-trait model. For example, when sample sizes are small the item parameters of the 3-PAR model can be poorly determined, and the researcher may be better off with the 1-PAR model or with no latent-trait model. Also, it is theoretically possible to select items and construct tests that are consistent with 2-PAR or 1-PAR

models. In choosing a latent-trait model the researcher needs to keep in mind the uses to which the trait or item parameters estimates will be put. The use of an inappropriate model can introduce different types of errors in different situations.

References

Allen, M. J., & Yen, W. M. Introduction to measurement theory. Monterey, CA: Brooks/Cole, 1979.

California Tests of Basic Skills, Form S. Monterey, Ca: CTB/McGraw-Hill, 1973.

IMSL Library (7th ed.). Houston, TX: International Mathematical and Statistical Libraries, 1979.

Wood, R. L., Wingersky, M. S., & Lord, R. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM-76-6). Princeton: Educational Testing Service, 1976.

Table 1

Mean Item Chi-Squares
for the Simulation Data

Estimating Model	Generating Model		
	3-PAR	2-PAR	1-PAR
3-PAR	7.8	8.9	8.7
2-PAR	9.9	8.8	8.4
1-PAR	35.7	30.0	10.1

Table 2

Mean Item Chi-Squares
for the CTBS Tests

stimating Model	Read. Vocab.	Read. Comp.	Spell.	Lang. Mech.	Lang. Exp.	Math Comput.	Math Concepts	Math Applic.
3-PAR	11.9	11.4	8.5	14.5	9.5	12.5	11.3	9.8
2-PAR	14.4	13.0	8.9	12.1	10.6	14.7	12.6	12.1
1-PAR	30.7	32.4	24.3	27.2	29.4	26.3	24.6	25.8

Table 3

Correlations Between Estimated Item Discriminating Powers and Estimated Item Difficulties for the CTBS Tests

Estimating Model	Read. Vocab.	Read. Comp.	Spell.	Lang. Mech.	Lang. Exp.	Math Comput.	Math Concepts	Math Applic.
3-PAR	.53	.33	-.22	.56	-.18	.38	.41	.40
2-PAR	-.33	-.41	-.57	.23	-.47	.23	-.44	-.45

$$P_i(\theta_k) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_k - b_i)}}. \quad (1)$$

$$\chi^2 = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(O_{ij} - E_{ij})}, \quad (2)$$

$$E_{ij} = \frac{1}{N_j} \sum_{\substack{k \in \\ \text{cell } j}} \hat{P}_i(\hat{\theta}_k). \quad (3)$$

Figure 1

Relationships between estimated trait values

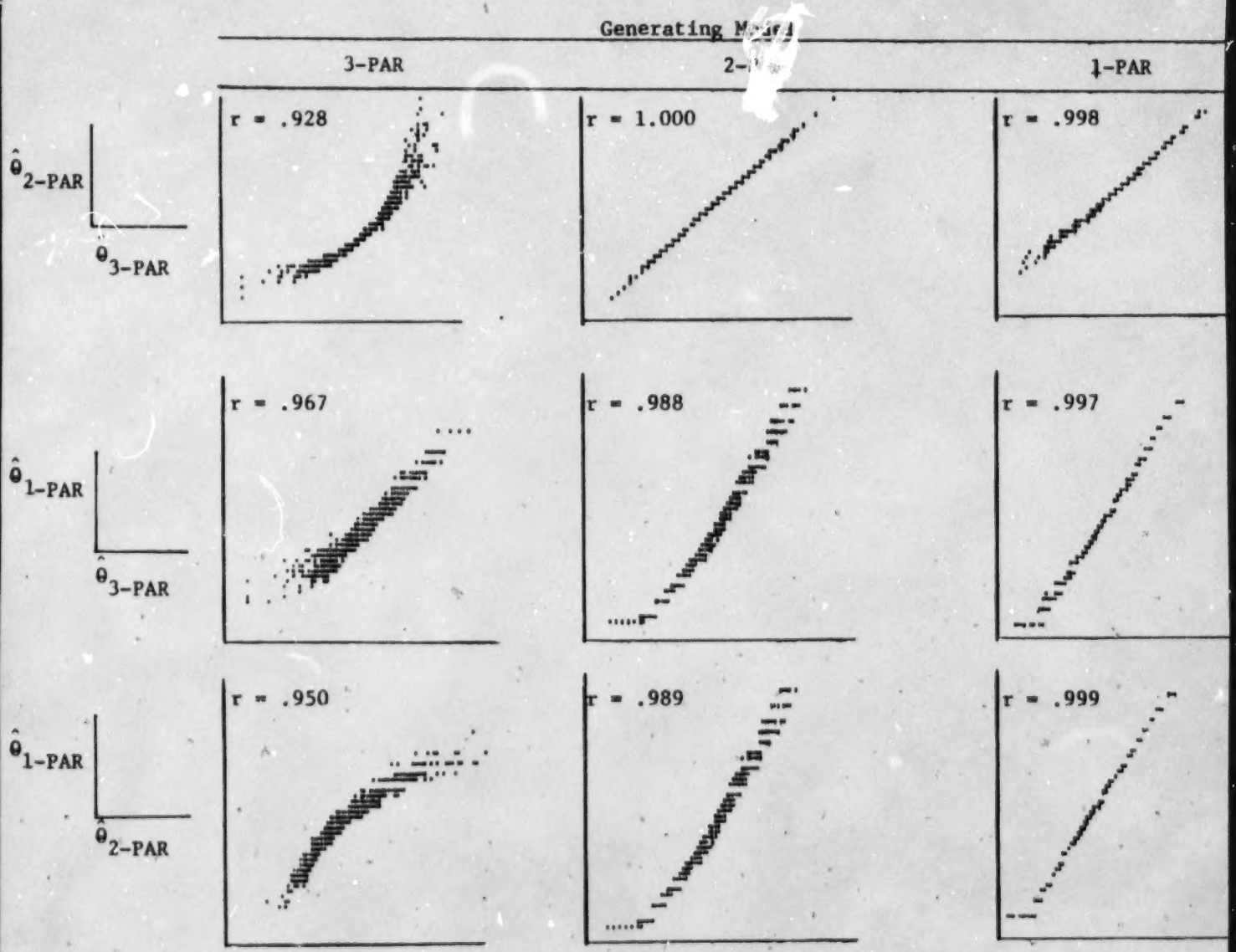

Figure 2

Relationships between estimated item difficulties

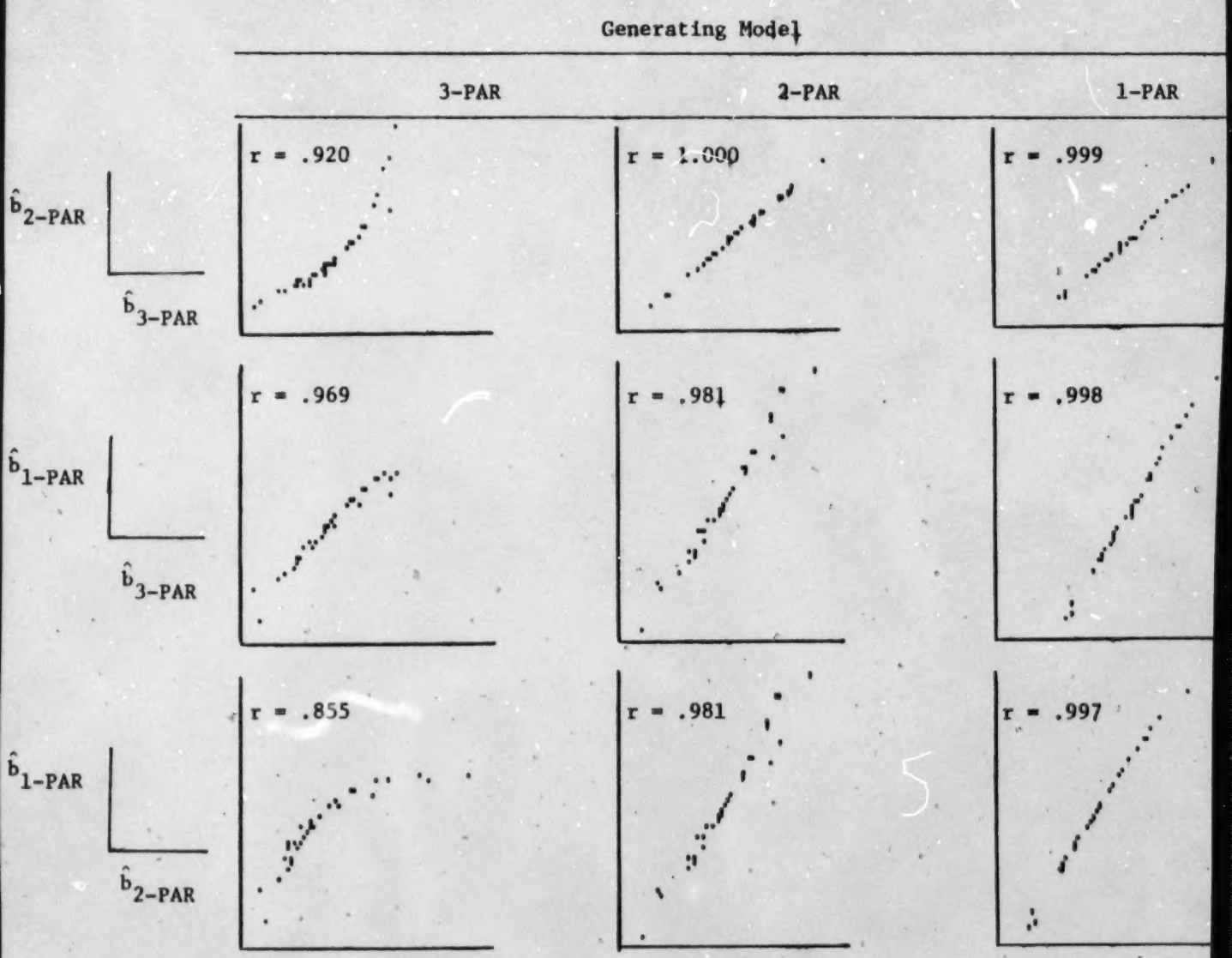

Figure 3

Relationships between estimated item discriminating powers across different models and between estimated item discriminating powers and item difficulties within one model

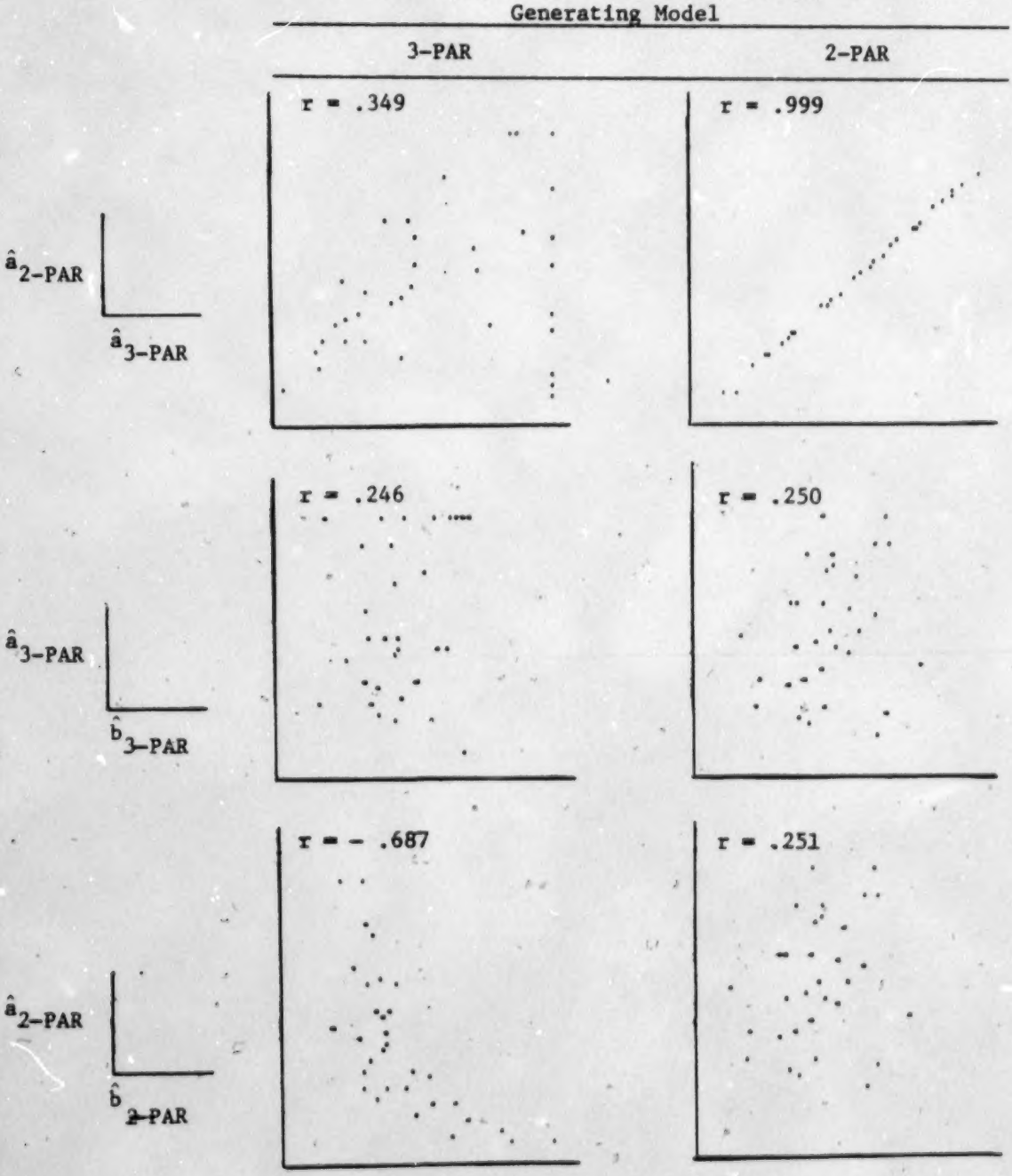

Figure 4

Relationships between trait estimates and between item parameter estimates for the Reading Vocabulary data